Supplementary Note for

Annotation-free quantification of RNA splicing using LeafCutter

Contents

List of Figures

1 Identifying alternatively spliced introns using LeafCutter

Starting from alignment files in `.bam` format, junctions from split-reads that map with minimum 6nt into each exon are extracted using a script we provide (1) based on two OLego helper scripts. Then, the LeafCutter clustering program (2) can be used to identify intron clusters supported by at least 30 (option `-m`) total reads (across all samples) and introns supported by more than 0.1% (option `-p`) of the total read counts for the entire cluster. The number of reads supporting each intron and cluster is then counted in all samples separately and collated in a table for downstream analyses (Supplementary Note Figure 1).

Because LeafCutter focuses on intron splicing rather than whole isoform quantification, alternative transcription start site or polyadenylation sites are not captured. However, several prevalent types of alternative splicing (Supplementary Figure 1) are equivalent to specific intron excision events.

2 Comparison of LeafCutter to other methods for differential splicing analysis

2.1 rMATS, MAJIQ, and Cufflinks2

To compare the ability of different software to detect differential splicing, a fair comparison needs to overcome (1) differences in p-value calibration, and (2) differences in what is being measured e.g. transcript ratios versus local splicing events. As test data, we chose to contrast lymphoblastoid cell lines (LCLs) derived from Yoruba individuals against LCLs derived from central european (CEU) individuals. We chose LCLs as they are homogeneous cell lines and splicing differences between populations should be subtle; both properties are favorable for comparing sensitivity of the methods.

To overcome (1) the problem of p-value calibration, we computed the empirical false discovery rate (FDR) as follows:

(a) First, we identify differential splicing between YRI and CEU LCLs using each method and record the p-value (1-posterior for MAJIQ, see subsection below) distribution for all tests.

(b) Next, we permuted labels on the samples such that $\sim 1/2$ of CEU samples are labeled as YRI samples and vice versa. We then run each method on these permuted samples and the p-value (1-posterior for MAJIQ) distribution are once again recorded.

(c) The number of differential splicing events discovered at a certain FDR (e.g. 5%) is defined as the maximal number of events with test p-value less than p in the real data (N_{real}) such that the number of events with test p-value less than p in the permuted data (N_{perm}) respects the following constraints $N_{\text{perm}}/(N_{\text{perm}} + N_{\text{real}}) < FDR$.

The resulting p-value distributions of the 3v3, 5v5, 10v10, and 15v15 comparisons are shown in (Supplementary Figure 7). We observed that LeafCutter p-values were generally well-calibrated, which resulted in the largest number of differentially spliced events compared to rMATS, MAJIQ, and Cufflinks2.

We observed that Cufflinks2 p-values were very conservative (see Cufflinks subsection below). We therefore report the number of significantly differentially spliced events from Cufflinks2 directly. Interestingly, Cufflinks2 reports 19 significantly different splicing events in the 3v3 comparison, but not in comparisons with large sample sizes.

To overcome (2) the problem of differences in what events are being measured, we collapsed all events in rMATS and MAJIQ that shared a single splice site into a single event (as is done in LeafCutter).

2.2 MAJIQ

Instead of computing p-values for differentially splicing tests, MAJIQ computes posterior values reflecting the confidence that a splicing event is differentially spliced at a $\Delta\Psi$ of at least P which is an user-defined parameter. In our tests, we chose P to be 0.05. Choosing other values of P, e.g. 0.01 resulted in similar or worse performance.

In principle it should be possible to use the posterior probabilities from MAJIQ's Bayesian model to directly control FDR. In particular, taking events with posterior probabilities 1-F should control FDR at F. However, our permutation analysis shows this is clearly not the case since this approach results in a highly inflated false positive rate (FPR) under the null. The fact that MAJIQ does not seem to give "true" posterior probabilities suggests some degree of model mis-specification, i.e. that the statistics of real RNA-seq counts do not quite match the assumptions made by the MAJIQ differential splicing model.

2.3 Cufflinks2

We sought to understand the source of Cuffdiff2's overly conservative p-value distribution under the null. To test for differential isoform usage for a specific gene Cuffdiff2 considers estimated isoform usage proportions for a gene in two groups, denoted $\hat{\kappa}^A$ and $\hat{\kappa}^B$ as well as associated posterior covariances $\hat{\Sigma}^A$ and $\hat{\Sigma}^B$. The test statistic used is the Jensen-Shannon distance (JSD), $d = \sqrt{KL(\hat{\kappa}^A|m) + KL(\hat{\kappa}^B|m)}$ with $m = \frac{1}{2}\hat{\kappa}^A + \frac{1}{2}\hat{\kappa}^B$. Under the null $\hat{\kappa}^A$ and $\hat{\kappa}^B$ are drawn from the same distribution, which Cuffdiff2 assumes to be multivariate normal. To approximate the sampling distribution of d, 10^5 *pairs* of samples are drawn from $N(\hat{\kappa}^A, \hat{\Sigma}^A)$, and the JSD for each pair. The procedure is repeated using $N(\hat{\kappa}^B, \hat{\Sigma}^B)$ and the two resulting empirical p-values are averaged.

To test the calibration of this procedure we simulated an idealized scenario where 1000 reads in each of two conditions are unambiguously mapped to 5 isoforms of a gene. The true (shared) usage proportions are sampled uniformly from the 5-simplex. Per condition counts are sampled from a Dirichlet-multinomial distribution to model overdispersion, with a concentration parameter $c = 10$, typical for RNA-seq data. We obtained maximum likelihood estimates of $\hat{\kappa}^A$ and $\hat{\kappa}^B$ under the "best-case" scenario of knowing the true c, and corresponding $\hat{\Sigma}^A$ and $\hat{\Sigma}^B$ estimates using the inverse Hessian of the log likelihood function. We then performed the Cuffdiff2 procedure using these values. The whole procedure was repeated for 100 different simulated true usage proportions. This procedure recapitulates the overly conservative p-value distribution (Supplementary Note Figure 2 and Supplementary Figure 7) we observed when applying Cuffdiff2 to permuted real RNA-seq data. We hypothesize that the root cause of the problem is that the multivariate normal is a poor approximation for distributions constrained to the simplex, and as a result the estimated sampling distribution of d is considerably more dispersed than it should be.

2.4 Comparison of false negative rates

To evaluate the false negative rates of differential splicing methods, we simulated sequencing reads for 160 protein coding genes each with 2 to 15 transcripts. For each gene, we only considered transcripts that differed by at least one overlapping intron when compared to another transcript to avoid cases where two transcripts only differ e.g. in the first or last exon or in an intron retention event (neither of which LeafCutter aims to detect). We then simulated reads from these transcript models as follows:

1. We simulated 8 biological samples each with 5 technical replicates.

2. For each gene, we set a random transcript's expression to 1X (no change), 1X, 1X, 1.1X, 1.25X, 1.5X, 3X, and 5X in the 8 biological samples in random order (note that we set 1X for 3 of 8 samples, so there are 3 comparisons with no change of transcript expression; we used these to compute false positive rates).

3. We used `polyester`[1] to simulate sequencing reads, obtaining 8 × 5 = 40 RNA-seq samples (we used default parameters, e.g. 30X coverage and default error distributions).

4. We mapped reads from each sample using STAR and applied all four differential splicing detection methods on all pairwise (8 choose 2) = 28 comparisons.

5. We computed the effective transcript fold-change for each gene (a transcript might be set to 1.5X and 3X in the two samples that are being compared resulting in a effective fold change of 2X) in all 28 pairwise comparisons.

6. We then collected all p-values for every gene/comparison (min p-value/max posterior if more than one splicing event is tested per gene) and plotted their differential splicing test p-values binned by their effective transcript fold-change (Supplementary Note Figure 3).

7. For each effective transcript fold-change, we computed the true positive and false positive rates for all possible p-value or posterior cutoffs (Supplementary Figure 8).

From these simulations and the receiver operating characteristic (ROC) curves, we conclude that while Cufflinks2 appears to detect more transcripts with 1.1 fold-difference at reasonable FDRs, LeafCutter outperforms all three other methods when transcripts differed by 1.25-fold or more (Supplementary Figure 8). Of the four methods tested, Cufflinks2 is the only method that estimates transcript levels, which might explain its higher power in detecting small differences in transcript expression. Interestingly, the performance of MAJIQ and LeafCutter were nearly identical when evaluated on transcripts that differed by 3-fold or more, but LeafCutter outperformed MAJIQ when differences were more subtle. This can be explained by the observation that LeafCutter has a lower false positive rate than compared to MAJIQ (see LeafCutter and MAJIQ panels at 1X effective fold-change in Supplementary Figure 3).

2.5 Additional comparisons

As further comparisons and to ensure that the differentially spliced events detected using LeafCutter are not simply noise. We first asked about the correlation of p-values between comparisons with varying sample sizes. Here, we only compared LeafCutter to rMATS as MAJIQ do not report p-values and Cufflinks2 p-values are overly conservative. To do this, we computed the Spearman correlation of the $-\log p$ of the tested introns in the 15v15 comparison versus the corresponding $-\log p$ of the tested introns in the 3v3, 5v5 and 10v10 comparisons. As expected, the correlations increase monotonically for both methods as sample size increases reflecting an increase in precision in our effect size estimates (Supplementary Note Figure 4a). However, we do observe a significantly higher correlation for LeafCutter compared to rMATS, suggesting that LeafCutter is more robust to comparisons involving fewer samples.

We further observed that the ability of LeafCutter to recall genes with evidence of differentially splicing discovered using an rMATS analysis was similar to that of MAJIQ, while Cufflinks2 showed the worst performance of all (Supplementary Note Figure 4b).

To estimate the concordance between methods, we ranked genes by our differential splicing p-values, and asked about concordance at different bins of significance levels (50 genes per bin). We found that for the most significant bin (i.e. the top 50 most significantly differentially spliced genes), the concordance was high (65–75%) between rMATS and LeafCutter (we used a p-value cutoff of 0.05 to determine concordance) and even higher (80–82%) between LeafCutter or rMATS top genes and MAJIQ genes (we used a posterior > 0.99 to determine concordance). These observations (Supplementary Note Figure 5a,b) are in line with our expectation that concordance rates rapidly decrease as our power to detect differentially spliced genes drops to zero.

Because LeafCutter, rMATS, and MAJIQ all measure splicing at a local level and not at the gene/isoform level, we next verified how consistent LeafCutter was with other predictions in terms of these local events. To this end, we ranked LeafCutter associations in terms of their p-values and asked whether LeafCutter introns shared at least one splice site with introns that were predicted to be differentially spliced by rMATS ($p < 0.01$) and MAJIQ (posterior > 0.95). We found that $\sim 90\%$ of the introns that were found to be most significantly differentially spliced using LeafCutter shared a splice site with rMATS and MAJIQ, suggesting that LeafCutter identified the same differentially spliced events (Supplementary Note Figure 5c). In contrast, only $\sim 60\%$ of the events shared a spliced site when no associations was in LeafCutter ($p > 0.5$). Although 60% might appear high for the sharing between two "random" introns, it is useful to note that these are conditioned on introns that show (1) alternative splicing and (2) are differential spliced in rMATS or MAJIQ. The random overlap between LeafCutter-tested introns and rMATS-tested introns is less than 20%.

2.6 RAM usage

To measure RAM usage across methods, we used a custom script which calls `strace -e trace=mmap ,munmap,brk` on the main programs, except for rMATS. We found that rMATS launched additional processes that are not measured directly. We therefore ran our custom script on `rMATS.3.2.5/processGTF.BAMs.py` which appears to be the most RAM intensive script of the rMATS pipeline.

3 RNA-seq data processing

3.1 GTEx for intron discovery

We downloaded 2,192 RNA-seq samples from GTEx (Supplementary Note Table 1). To analyze these, we used OLego (v1.1.5)[2] to map the RNA-seq reads to the human genome (hg19) and processed the resulting `.bam` files using LeafCutter. Specifically, we used the following command:

```
olego -v -j hg19.intron.hmr.brainmicro.bed -e 6 hg19.fa.
```

The choice of OLego[2] is based on our previous experience that it performs well for discovering unannotated exons of small length (e.g. 9nt micro-exons)[3]. OLego is a program specifically designed for de novo spliced mapping of mRNA-seq reads, while STAR[4] does best when a set of junction is provided. Since a chief objective of our GTEx analysis was to identify novel exons and to identify conserved alternative splicing events across multiple species with annotations worse than that of human, we used OLego for our GTEx differential splicing analyses (we used STAR for sQTL analyses because of fast running time and high accuracy in mapping). To quantify the differences in mapping of the two aligners, we picked at random five RNA-seq samples from the GTEx consortium that we previously aligned using OLego and re-aligned them using STAR. We next analyzed the correspondence between the number of junction reads for each junction across the two aligners (Supplementary Note Figure 6). We found that while there are junctions whose read counts are orders of magnitude different, only 4.8% of junctions differed by a count fold-difference of 1.1 or more (0.94% of junctions differed by a count fold-difference of 2 or more).

3.2 GEUVADIS (YRI) for sQTL methods comparison

To compare splicing QTL (sQTL) calling methods, we aligned 85 YRI LCL samples from GEUVADIS using STAR two-pass and used WASP to remove reads that mapped with allelic biases[5]. These aligned reads were used as starting point for each of the sQTL calling methods. Specifically, the following command was used:

```
STAR --genomeDir STAR_index --twopassMode --outSAMstrandField intronMotif
--readFilesCommand zcat --outSAMtype BAM Unsorted
```

3.3 GEUVADIS (CEU) for sQTL mapping

To control for differences in mapping procedures, we downloaded the `.bam` files directly from ArrayExpress (E-GEUV-3) and processed them using LeafCutter to obtain intron clusters and quantifications. We recommend the use of WASP[5] to correct for biases caused by allelic reads. However, to make our comparison to other tools fair, we used the aligned reads available on ArrayExpress, and removed all clusters with an association to a SNP that overlap junction reads (see section entitled "sQTL mapping using LeafCutter"). This approach is conservative as some allelic reads do not map with a bias.

3.4 GTEx for sQTLs mapping

Again, to control for differences in mapping procedures, we used the `.bam` files provided by the GTEx consortium for sQTL mapping, and removed all clusters with an association to a SNP that overlap junction reads.

4 Identification of unannotated introns in tissues from GTEx

To obtain a comprehensive set of annotated introns, we downloaded the GENCODE (v19), UCSC, and RefSeq annotation databases in `.gtf` format. We classified introns as annotated if their 5' and 3' splice sites correspond to the end and start, respectively, of two consecutive exons in at least one transcript. As such it is possible that both 5' and 3' splices sites of a novel intron are annotated. We note that although a large proportion of annotated introns are present in all three databases, we found that the GENCODE annotation has the most comprehensive list of introns.

To estimate the number of unannotated alternatively excised (AE) introns, we first mapped 2,192 RNA-seq samples from 14 tissues (GTEx) to the human genome (hg19) using OLego, allowing *de novo* splice junction predictions. We then used LeafCutter to identify alternatively excised introns by pooling all junction reads. We then restricted our analyses to AE introns that were supported by at least 20% of the total number of reads that support introns from the clusters they belong to in at least 25% of all samples, considering each tissue separately. Although there is no minimum read count (an intron supported by 20 reads, 20% of 100, is less likely to be the outcome of noisy splicing than one supported by 2 reads out of 10), we reasoned that requiring 20% percent-splicing in 25% of all samples will filter out most sequencing technical artifact and noisy splicing. Importantly, using different cutoffs does not alter qualitatively our conclusions. This resulted in 70,722 AE introns that met these criteria, of which 22,278 (31.5%) AE introns were absent from all three annotation databases.

To investigate the functionality of these unannotated introns, we asked whether the unannotated splice sites of the 22,278 AE novel introns show signature of sequence conservation across vertebrates. To do this, we divided splice sites into three classes: (1) control splice sites, which are annotated in one or more databases, but whose cognate splice site is unannotated, (2) the cognate splice site itself, and (3) splice sites of introns, for which both splice sites are unannotated. To compute sequence conservation, we average the phastCons score of the predicted splice sites (over 96% of which are AG/GT) plus 2 flanking bases. Interestingly, we find that the average sequence conservation of unannotated splice sites is higher if its cognate splice site is annotated (Figure 2a, Supplementary Figure 4).

4.1 Validation of unannotated junctions in Intropolis

To verify that these unannotated splicing events are not a result of mapping errors or artefact unique to samples from GTEx, we examined the number of splicing junctions that could also be found in the Short Read Archive (SRA) using `Intropolis`[6] (note that GTEx samples were excluded from the SRA). `Intropolis` processed 21,504 human RNA-seq samples from the Sequence Read Archive (SRA) using RAIL-RNA to align and refine junction calls to improve sensitivity[7]. This analysis therefore provides an additional replication of the RNA-seq aligner (i.e. OLego) that we used to identify unannotated splicing events. Because the SRA does not collect uniform cell-type or tissue labels for each sample, we used the cell-type or tissue labels predicted by `phenopredict`[8] to assign tissue identity to each SRA sample. Using this data, we quantified the number of alternatively spliced junctions identified in our study that can also be found in SRA samples (Supplementary Figure 2). Overall, we found that, for instance, 86% of all novel junctions identified in GTEx testis using LeafCutter could be replicated in testis samples from the SRA (94% of unannotated heart junctions could be found in heart SRA samples). This is particularly impressive because (1) at most 56% of all unannotated junctions could be found in any other SRA tissues and (2) considering all tissues together increased the proportion of unannotated junctions “replicated” by only 4%, to 90%. These observations cannot be simply explained by a better sampling of testis in the SRA, as, for example, only 77% of the novel heart junctions could be found in SRA testis samples versus 94% in SRA heart samples.

Because the analysis above only quantified presence or absence of the unannotated junctions in at least one sample from each tissue, we next characterized unannotated junctions by examining the proportion of samples in which they could be found by tissue (Supplementary Note Figure 7). As expected, we found that unannotated junctions discovered in a given tissue tend to be present in a significant higher proportion of samples from the same, corresponding, tissue. Again, this suggests that unannotated junctions likely represent real splicing events that were not previously annotated as they tend to be highly tissue-specific.

To further profile this set of unannotated introns, we quantified their tissue-specificity, their levels of usage, and the type of splicing patterns they generally correspond to. As expected, we found that the vast majority of novel junctions were present in only a single GTEx tissue (Supplementary Figure 3a). Similarly, we found that novel junctions identified in a tissue were used at a significantly higher levels in the corresponding tissue than in other tissues (Supplementary Figure 3b), the differences were particularly striking for novel junctions discovered in testis. When we characterized the type of splicing events in which the unannotated introns were apart of, we found that, interestingly, 31.7% of all clusters with unannotated introns were complex, i.e. included at least one exon skipping and one alternative splice site event. This is nearly twice as many as compared to the 16.6% of complex clusters that are annotated. Overall, we conclude that unannotated junctions are relatively lowly used, tend to be tissue-specific, and often involve complex splicing patterns.

5 Statistical models

For cluster C containing J possible introns, let $\mathbf{y}_{ij}$ denote the count for sample i and intron j (and cluster total $\mathbf{n}_{iC} = \sum_{j'} \mathbf{y}_{ij'}$), and $\mathbf{x}_i$ denote a P-vector of covariates.

5.1 Beta-binomial GLM.

Our initial approach was to test each specific intron j of a cluster using

$$\mathbf{y}_{ij}|\mathbf{n}_{iC} \sim \mathcal{BB}(\mathbf{n}_{iC}, \alpha p_i, \alpha(1-p_i)), \tag{1}$$

$$p_i = \sigma(\mathbf{x}_i \beta + \mu) \tag{2}$$

where $\mathcal{BB}$ is the beta-binomial distribution and $\sigma(x) = 1/(1+e^{-x})$ is the logistic function. Here the parameters to be learnt are the P-vector β, intercept μ and concentration parameter α. Higher values of α correspond to the underlying beta distribution concentrating around p_i, and therefore to less count overdispersion. In particular as $\alpha \to \infty$ the $\mathcal{BB}$ likelihood converges to a multinomial likelihood, recovering a logistic regression model.

Optimization. For both the beta-binomial and Dirichlet-multinomial models we use the Bayesian probabilistic programming language Stan[9] to define the model, generate efficient C++ code for likelihood and gradient calculation, and to perform optimization using LBFGS.

Regularization. For some cases the likelihood as a function of the overdispersion parameter can be extremely flat, leading to numerical instability. In order to stabilize the optimization we use very weak regularization in the form of the prior

$$\alpha \sim \text{Gamma}(1 + 10^{-4}, 10^{-4}) \tag{3}$$

We experimented with two different versions of the $\mathcal{DM}$ GLM. The first uses a shared concentration parameter $\alpha_j = \alpha$ for all introns j in a cluster (the beta-binomial GLM is a special case of this model). The second allows a different α_j for each intron in the cluster.

Identifiability. The $\mathcal{DM}$ GLM shares with the more standard Multinomial GLM that has a spurious degree of freedom: in particular, adding a constant to the input of the softmax does not change its output. To remove this degree of freedom from the model we parameterize each β_j as

$$\beta_{jp} := \bar{\beta}_p(\tilde{\beta}_{jp} - \frac{1}{J}) \tag{4}$$

where $\tilde{\beta}_{1p}, ..., \tilde{\beta}_{Jp}$ is constrained to lie on the J-simplex, i.e. $\tilde{\beta}_{jp} \geq 0, \sum_j \tilde{\beta}_{jp} = 1$, a constraint Stan naturally handles using a change of variables.

5.2 Likelihood ratio tests

Likelihood ratio tests are generally better calibrated than alternatives such as Wald statistics for testing for the significance of covariates, especially for modest sample sizes. We optimize wrt to β, μ, α separately for the null and alternative models (excluding and including the group indicator x respectively) to obtain log likelihoods λ_0 and λ_1 (for efficiency we initialize the optimization for the alternative model using the null model parameters) and then perform a likelihood ratio test: under the null $2(\lambda_1 - \lambda_0) \sim \chi^2_\rho$ where ρ is the appropriate degrees of freedom. For the beta-binomial GLM $\rho = P_1 - P_0$ where P_0 and P_1 are the number of covariates in the null and alternative models respectively. For the Dirichlet-multinomial GLM we have $\rho = (J - 1)(P_1 - P_0)$ where J is the number of introns in the cluster.

6 Differential intron excision analyses

6.1 Identification of tissue-dependent intron excision levels

We used LeafCutter's Dirichlet-multinomial GLM to identify intron clusters with at least one differentially excised intron. We searched for intron excision level differences between all tissue pairs. However, we should note that owing to sample size differences, we will have different power to detect differential splicing of varying magnitude between pairs (we can detect splicing differences of small magnitude only in comparisons with large sample sizes). When we hierarchically clustered all samples according to the intron excision levels of introns that were present (i.e. were supported by reads) in all species, we saw a mix between tissue and species clustering (Supplementary Figure 10). However, when we conditioned on introns that were differentially excised across human tissue pairs according to LeafCutter, we saw a clear clustering by tissue (Figure 3d).

6.2 Effectiveness at small sample sizes

RNA-seq experiments are often performed on a handful of samples only. To determine whether LeafCutter is effective in this setting we performed clustering, quantification and differential intron usage analysis on 4 male brain and 4 male muscle samples from GTEx. As a "bronze standard" we additionally performed quantification and differential splicing on 110 muscle and 110 brain samples (using the introns and clusters identified using 8 samples). With only $N = 8$ samples, LeafCutter appears to be well-calibrated under permutations (Supplementary Figure 9a) and has sufficient power to detect 885 clusters with evidence of differential intron usage (FDR 10%, maximum absolute effect size > 1.5), compared to 1906 found at $N = 220$. The per cluster p-values are highly correlated between the small and full sample sizes ($R^2 = 0.72$, Supplementary Figure 9b), and 98% of the clusters significant at $N = 8$ are also significant at $N = 220$. Per intron effect sizes between the two sample size settings are also highly correlated ($R^2 = 0.49$, Supplementary Figure 9c), although as expected the variance of the $N = 8$ effect sizes is large. This is particularly the case when the intron is only observed at all in one of the two tissues (Supplementary Figure 9d).

6.3 Pan-mammalian tissue clustering of intron excision profiles

To evaluate the conservation of intron excision profiles across mammalian tissues, we used OLego to map RNA-seq data[10] from eight organs (testes, heart, kidney, liver, lung, brain, colon, and spleen) in four mammals (mouse, rat, cow, and rhesus macaque) to their respective genomes. We then projected all introns supported by RNA-seq reads onto the human genome using liftOver and clustered projected introns from all four mammals and human GTEx samples using LeafCutter. We then focused on four disjoint pairwise comparisons (Testis vs Kidney, Muscle vs Colon, Heart vs Lung, and Brain vs Liver, Figure 11).

7 sQTL mapping using LeafCutter

7.1 Mapping sQTLs in GEUVADIS LCL samples (linear regression)

To map sQTLs in GEUVADIS LCLs samples, we restricted our analysis on 372 samples derived from European individuals. We downloaded genotype files from ArrayExpress (E-GEUV-1). We used LeafCutter to obtain read proportions for all introns within alternatively excised intron clusters. We then standardized the values across individuals for each intron and quantile normalized across introns[11] and used this as our phenotype matrix. We then used linear regression (as implemented in fastqtl)[12] to test for associations between variants (MAF ≥ 0.05) within 100kb of intron clusters and the rows of our phenotype matrix that correspond to the introns within each cluster. As covariate, we used the first 3 principal components of the genotype matrix plus the first 15 principal components of the phenotype matrix. To estimate the number of sQTLs at any given false discovery rate (FDR), we used the correct p-values from fastqtl, and then used Bonferroni correction to control for the number of introns we test per cluster (note that this is conservative). We then use Benjamini-Hochberg to estimate the FDR (sample permutations show that our association p-values at this step are well calibrated).

Unlike for YRI RNA-seq data where we used WASP[5] to correct for biases in allelic reads, we did not correct for biases caused by allelic reads for the CEU comparisons to keep comparisons fair with previous GEUVADIS analyses. To avoid biases, we removed all associations that might be caused by SNPs that overlap junction reads. To do this, we removed all intron clusters that had a variant that were 70 or fewer base pairs (GEUVADIS RNA-seq read length is 75bp and at least 6nt must overlap with all exons) away from the splice sites (in the exonic part).

7.2 sQTL mapping comparison between LeafCutter, Cufflinks2 and Altrans

We ran LeafCutter, Cufflinks2 and Altrans to estimate isoform and splicing events usage, respectively, on all 85 Yoruba WASP-processed[5] RNA-seq aligned data. We then standardized the values across individuals for each isoform/splicing event usage and quantile normalized across introns[11]. As covariate, we used the first 3 principal components of the genotype matrix plus the first 15 principal components of the phenotype matrix. We then used fastqtl[12] to test for associations between variants (MAF ≥ 0.05) within 50kb of the transcript (Cufflinks), the splicing event (Altrans), or splicing cluster (LeafCutter). To estimate the number of sQTLs at any given false discovery rate (FDR), we used the correct p-values from fastqtl, then use Benjamini-Hochberg to estimate the FDR. Altrans discovers splicing events using a forward and a reverse pass on the aligned RNA-seq data, thereby producing two measurement tables. To allow fair comparison between Altrans and the other methods, we combined forward and reverse splicing QTLs and collapsed all events that shared a splice site together (as is done in LeafCutter).

7.3 Sharing of sQTL discoveries between LeafCutter, Cufflinks2 and Altrans

To quantify the proportion of LCL sQTLs that are shared between Cufflinks2, Altrans, and LeafCutter, we first took the most significant SNP-gene/cluster pairs for every gene/clusters that had a sQTL at a 10% FDR. Note here that the following observations were qualitatively the same when we used a 1% FDR cutoff. We then collected the p-values of the associations of the SNP-gene pairs (when there were more than one splicing event tested per genes, we took the minimum p-values times the number of tested events) (Supplementary Note Figure 8) and used the Storey's π_0 method[13] to estimate the proportion of shared discoveries (Supplementary Figure 12).

Overall, we find a higher pairwise sQTL sharing between LeafCutter and either of the two other methods (Altrans and Cufflinks2) than compared to the sharing between Altrans and Cufflinks2. Conversely, we found that while LeafCutter identified more sQTLs at 10% FDR, LeafCutter sQTLs were more enriched in low sQTL p-values as measured by Altrans or Cufflinks2. These observations suggest that LeafCutter is both more sensitive (lower proportion of false negatives) and more accurate (lower proportion of false positives).

7.4 Mapping sQTLs in GEUVADIS LCL samples (Dirichlet-multinomial GLM)

In addition to using linear regression, we also used LeafCutter's Dirichlet-multinomial GLM to map sQTLs. This approach has two main advantages: (1) it accounts for the over-dispersion of read count data, and (2) it combines signal from changes in intron excision levels across the entire cluster instead of considering each intron independently. However, when we applied to our GEUVADIS data and controlled FDR using permutations, we found fewer sQTLs than our linear model approach, likely driven by clusters with heavy-tailed count distributions which are effectively handled by the quantile normalization in the linear approach.

7.5 Mapping sQTLs in four GTEx tissues

To identify sQTLs in GTEx tissues, we used the same strategy as in GEUVADIS LCLs (linear regression). However, we used the first 5 genotype PCs and the first 10 PCs as covariates (5+10 instead of 3+15).

8 sQTL analyses

8.1 Identification of functional enrichment of sQTLs

As in earlier work[14], we found that LeafCutter sQTLs were strongly enriched within or close to the cluster they affect (Supplementary Figure 13). To identify functional categories enriched in sQTLs, we first annotated all variants using `SnpEff` version 4.1f. We next sampled at random ∼200,000 SNPs that are located near genes (i.e. had the annotation "Upstream", "Downstream", "Intronic", or were exonic variants). This is because we only test SNPs that are near genes. The number of sampled SNPs corresponds to 50 times the number of sQTLs identified in our study. We computed the log-fold enrichment in functional annotations of the top most significant sQTLs ($n = 4,543$) over this random sample of SNPs. Finally, to obtain confidence intervals, we repeated the random sampling procedure 500 times.

8.2 Comparison with GEUVADIS exon eQTLs, and trQTLs

Although LeafCutter does not explicitly search for genetic variants that are associated with differences in exon level splicing or transcript ratios, we expected that these variants will also affect intron excision, which are detected by LeafCutter. To verify this, we compared the distribution of p-values from the association between LeafCutter intron excision and genome-wide SNPs to the p-values from the association between LeafCutter intron excision and SNPs that were previously classified as exon eQTLs and transcription ratio QTLs in GEUVADIS. More specifically, we downloaded the list of exon eQTLs and trQTLs from ArrayExpress (E-GEUV-3) and for each exon/gene took the SNP with the strongest association to exon level or transcript ratio. We then computed the association p-values of these SNPs with all tested LeafCutter intron excision levels, using Bonferroni correction to adjust our p-values. As expected both exon eQTL and trQTL SNPs were enriched in strong associations to intron excision levels compared to random SNPs, and trQTL SNPs were most enriched in strong associations.

We next wished to verify that trQTLs detected in GEUVADIS were mostly identified as LeafCutter intron sQTLs. We again took the best trQTL SNP for each gene, and estimated the number that were associated with a cluster at a corrected p-value < 0.05. To correct for SNPs tested against multiple clusters, we used Bonferroni correction to adjust the p-value of the strongest association. We find that 399 (81.3%) of the 491 top trQTLs we tested are significantly associated ($p < 0.05$); this percentage is likely higher because our Bonferroni correction is conservative. Furthermore, as expected, when we use the same procedure to ask how many of the top 491 trQTLs are significantly associated to intron splicing when our sample labels are permuted, we find that only 4.7% are (our statistical tests are well calibrated; ∼5% of our tests should achieve a 0.05 significance under the null model).

8.3 Relationship between gene expression levels and power to detect sQTLs

We examined the expression profiles of the genes with significant sQTLs detected by LeafCutter. As expected we found a strong positive relationship between our power to detect a sQTL for a gene and the expression level of a gene (Supplementary Note Figure 9a). Indeed, while most annotated genes (including non protein-coding genes) were expressed at very low levels, we found almost no sQTLs for genes whose expression were less than 0.025 RPKM. While there is a clear decrease in LeafCutter's ability to identify sQTLs in lowly expressed genes (Supplementary Note Figure 9a), we were able to find sQTLs for many lowly-expressed genes, starting from 0.1 RPKM (Supplementary Note Figure 9b).

8.4 Replication of sQTLs across GTEx tissues

To estimate the proportion of sQTLs that are replicable across tissue types, we took the best SNP of each sQTL-cluster pair for each tissue and asked whether the sQTL association was significant ($p < 0.05$) in another tissue. This estimate is likely to be conservative as it does not account for incomplete power. The replication is therefore likely to be even higher than our current estimates of 75–93%.

8.5 Tissue-specific sQTLs

To identify tissue-specific sQTLs, we searched for genetic variants that were associated significantly with intron excision levels in one tissue, but not in any of the other three tissues ($p > 0.1$), requiring all tissues to have junction reads in the intron cluster.

9 LeafCutter sQTL signals in genome-wide association studies

To verify that LeafCutter sQTLs can help us identify disease-associated variants that function by modulating splicing, we downloaded summary statistics from two autoimmune GWAS studies (multiple sclerosis[15] and rheumatoid arthritis[16]) and looked for enrichment of strong association p-values among the top LeafCutter sQTLs and GEUVADIS gene eQTLs (we removed the extended MHC region from this analysis). We found that 1,205 LeafCutter sQTL SNPs and 901 GEUVADIS eQTL SNPs (the SNP with most significant p-value) were also tested (with >5% MAF) in the multiple sclerosis genome-wide association study, and that 3,069 LeafCutter sQTL SNPs and 2,250 GEUVADIS eQTL SNPs were tested in the rheumatoid arthritis study. We then took the QTLs and plotted the distribution of $-\log_{10}(p\text{-value})$ of their association to each trait separately. As expected[14], we found that LeafCutter sQTLs were more highly enriched in associations with low p-values compared to GEUVADIS eQTLs in multiple sclerosis and were similarly enriched in rheumatoid arthritis. This is notable because we considered a larger number of LeafCutter sQTLs than GEUVADIS eQTLs for both diseases. These observations suggest that LeafCutter allows us to identify as many or more disease-associated variants that *act* by affecting splicing as compared to those that *act* by affecting total expression levels.

9.1 Prediction Models and S-PrediXcan

Prediction models were trained by fitting Elastic-Net linear models to each gene for the expression models and to each intron cluster for the splicing models using nearby SNPs dosages as features. Before fitting the models, we removed non biallelic SNPs and any ambiguously stranded SNPs from the genotype data. We downloaded normalized and PEER corrected expression data from the GEUVADIS study. Intron excision traits were corrected for genetic principal components and covariates (as outlined above). Once the data had been preprocessed, for each gene or intron cluster, SNPs within 1Mb upstream and 1Mb downstream of their start and end sites were selected as variables for the model. Using the R package glmnet we fit a 10-fold cross-validated Elastic-Net linear model using a mixing parameter of 0.5 for each gene and intron cluster. Further details can be found in[17,18,19] and training pipelines can be downloaded from github.com/hakyimlab/PredictDBPipeline.

A total of 4625 gene associations were obtained for the genetic expression model, and 41196 intron quantification cluster associations for the splicing model, that had a model prediction FDR $< 5\%$ (computed from the correlation between cross validated prediction and observed values).

We downloaded genomewide association meta analysis (GWAMA) results for 40 phenotypes from 18 consortia and performed S-PrediXcan analysis using both expression and intron models. The full list of traits and consortia is displayed in Supplementary Note Table 3.

10 Processed data availability

See Supplementary Note Tables 3 and 4.

11 Supplementary Note Tables

Tissue	Sample Number
Heart	153
Testis	67
Spleen	7
Skin	340
Brain	422
Colon	86
Blood	270
Pancreas	66
Adipose Tissue	172
Lung	151
Esophagus	238
Muscle	176
Kidney	8
Liver	35

Supplementary Note Table 1: Sample sizes of processed GTEx RNA-seq short read data by tissue type.

Tissue	Number of individuals
Heart	95
Blood	170
Lung	128
Thyroid	118

Supplementary Note Table 2: Sample sizes of processed GTEx `.bam` files for sQTL mapping.

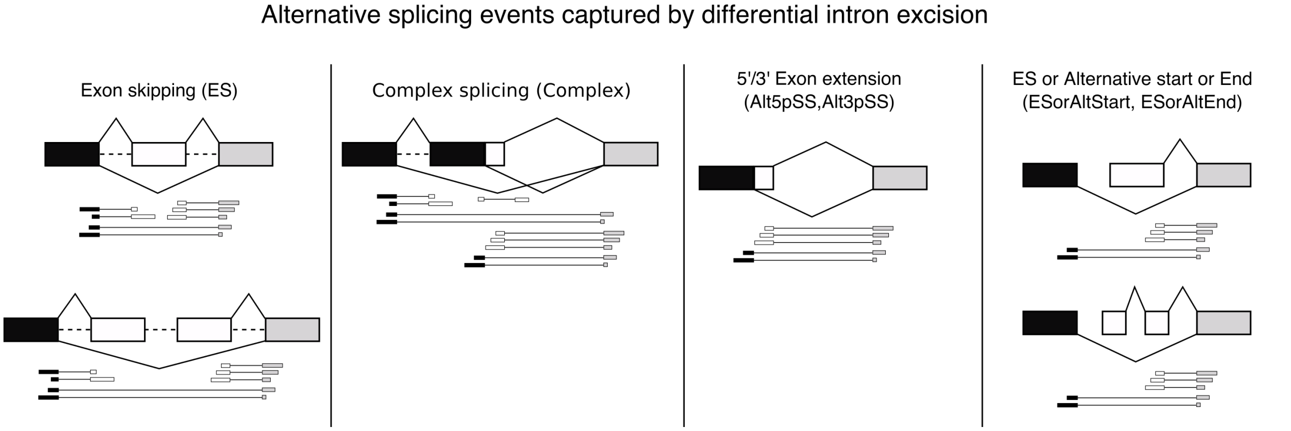
Consortium	Phenotype	URL
PGC	Attention Deficit/Hyperactivity Disorder	med.unc.edu/pgc/results-and-downloads
PGC	Bipolar Disorder	med.unc.edu/pgc/results-and-downloads
PGC	Major Depressive Disorder	med.unc.edu/pgc/results-and-downloads
PGC	Autistic Spectrum Disorder	med.unc.edu/pgc/results-and-downloads
PGC	Schizophrenia	med.unc.edu/pgc/results-and-downloads
CIAC	Clozapine-Induced Agranulocytosis	med.unc.edu/pgc/results-and-downloads
CONVERGE	Major Depressive Disorder	well.ox.ac.uk/converge
IGAP	Alzheimer	web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php
TAG	Tobacco Cigarettes per Day	med.unc.edu/pgc/results-and-downloads
IBD	Inflammatory Bowel Disease	ibdgenetics.org/
IBD	Ulcerative Colitis	ibdgenetics.org/
IBD	Crohn's Disease	ibdgenetics.org/
GIANT	Body Mass Index	broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
GIANT	Waist-to-Hip Ratio	broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
GIANT	Waist Circumference	broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
GIANT	Hip Circumference	broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files

Supplementary Note Table 3: List of Genome-wide Association Meta Analysis (GWAMA) Consortia and phenotypes.

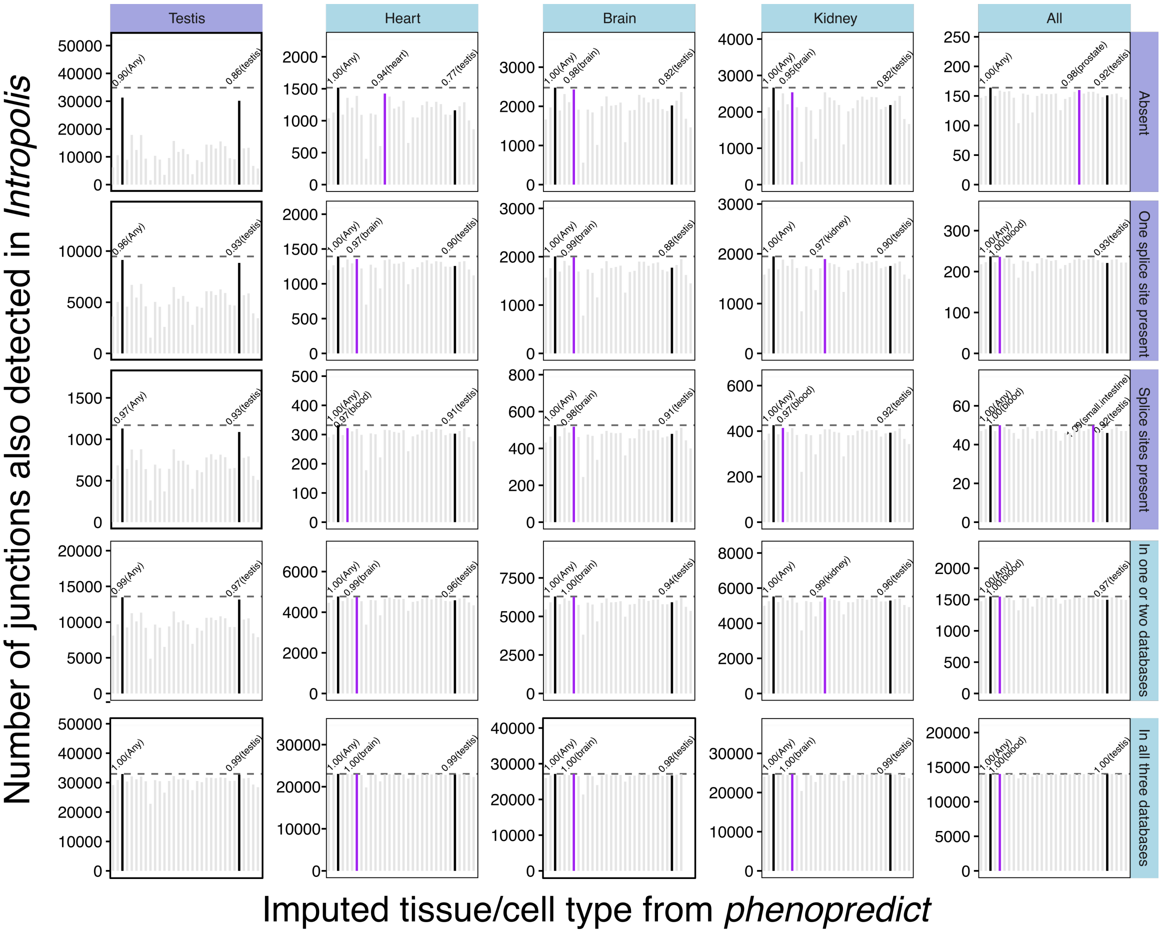
Data	Accession
RNA-seq and genotype (GEUVADIS)	E-GEUV-3 (ArrayExpress)
RNA-seq (Merkin et al., 2012)	GSE41637 (GEO)
RNA-seq and genotype (GTEx)	phs000424.v6.p1 (dbGaP)

Supplementary Note Table 4: RNA-seq accession codes.

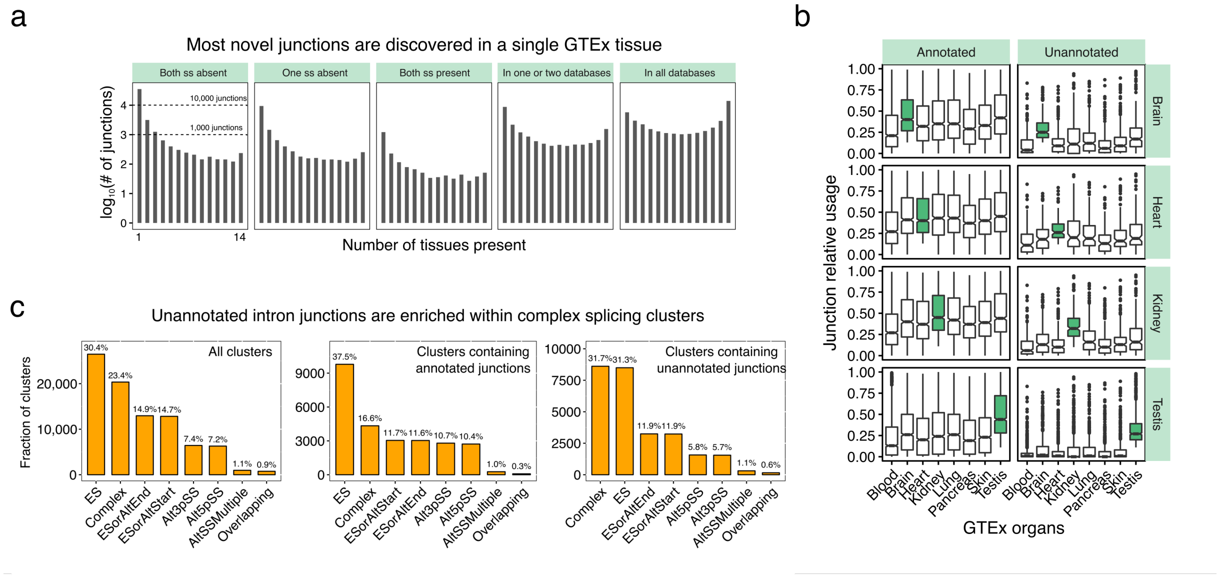
12 Supplementary Figures


Alternative splicing events captured by differential intron excision
Exon skipping (ES)
Complex splicing (Complex)
5'/3' Exon extension
(Alt5pSS,Alt3pSS)
ES or Alternative start or End
(ESorAltStart, ESorAltEnd)

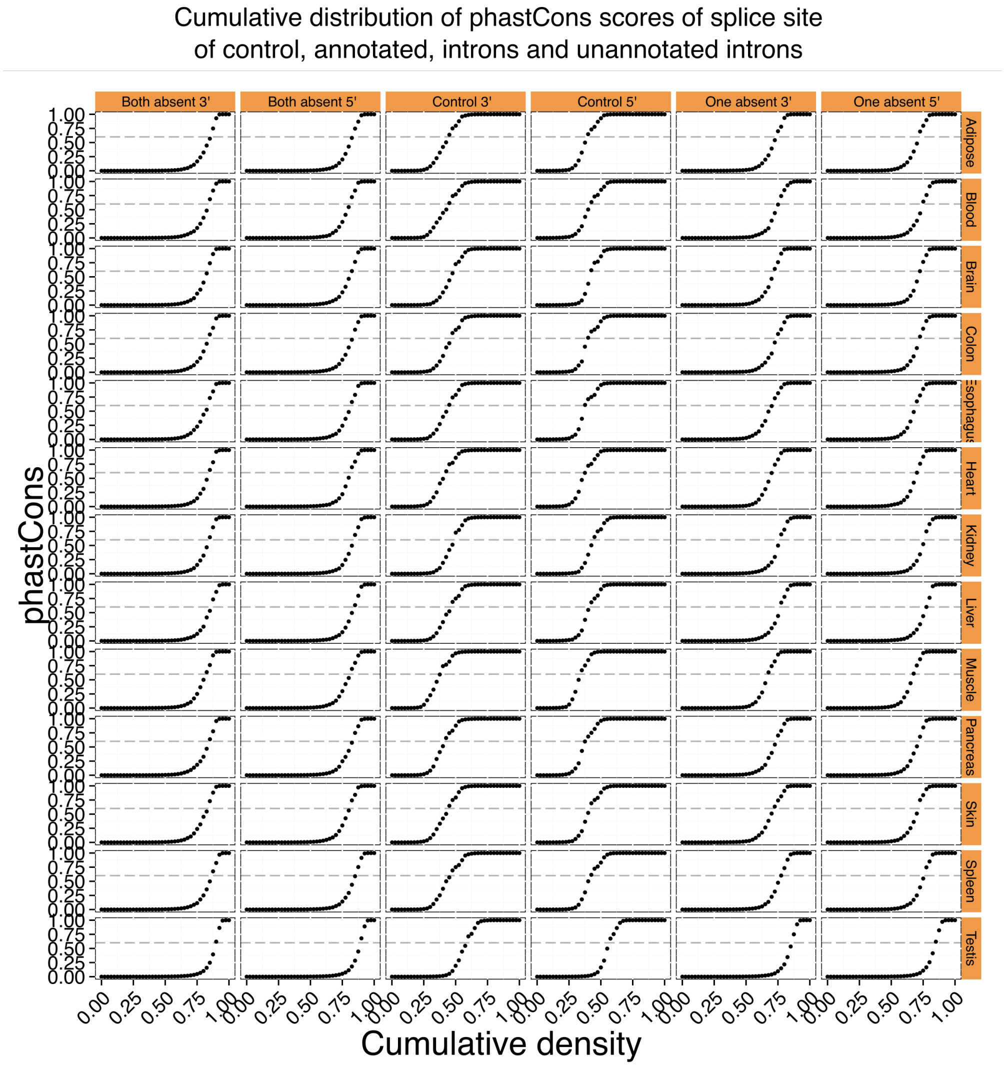
Supplementary Figure 1: Several types of common alternatively splicing events are captured by the alternative excision of introns.



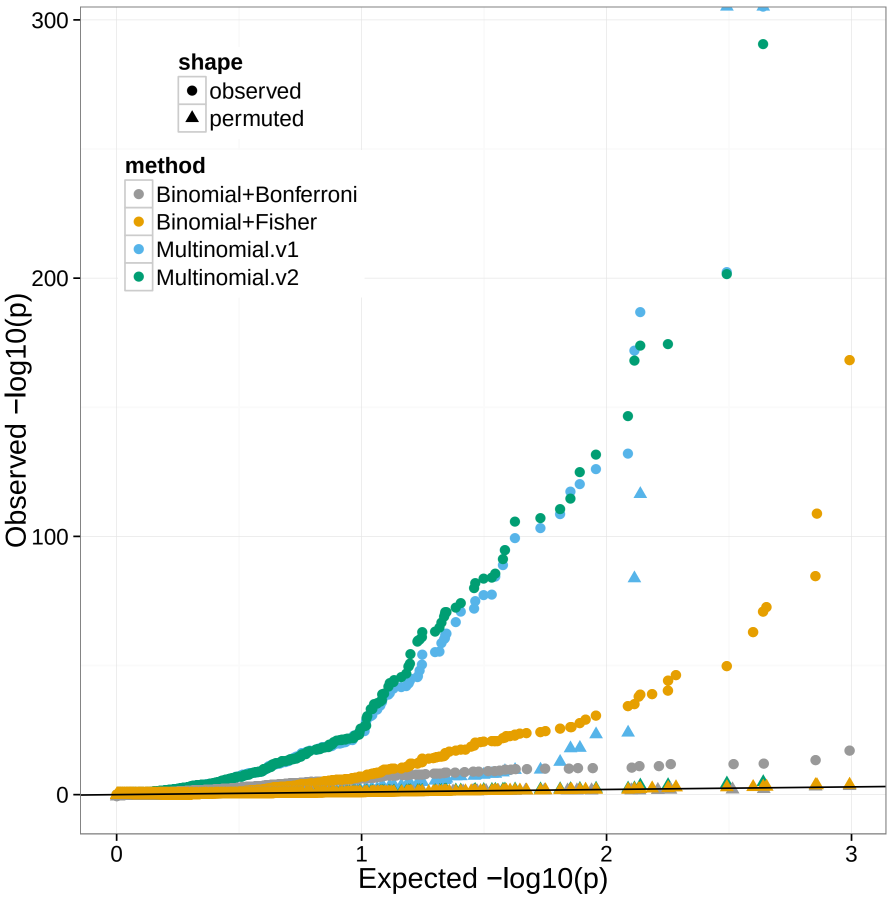
Supplementary Figure 2: Barplots showing the number of alternatively used junctions annotated from our GTEx analyses that were found in Intropolis[6]. phenopredict[8] was used to predict the tissue type corresponding to the SRA samples analyzed in Intropolis. For each set of junctions, the proportion of junctions that were found (at least 1 read) in any SRA sample (Any), or found in samples which were predicted to be from testis (Testis) are highlighted. The predicted tissues with the highest number of supported junctions are colored in purple. Eighty-six percent of all novel alternatively used testis junctions from our LeafCutter analysis could be found in testis samples within SRA (not including GTEx).



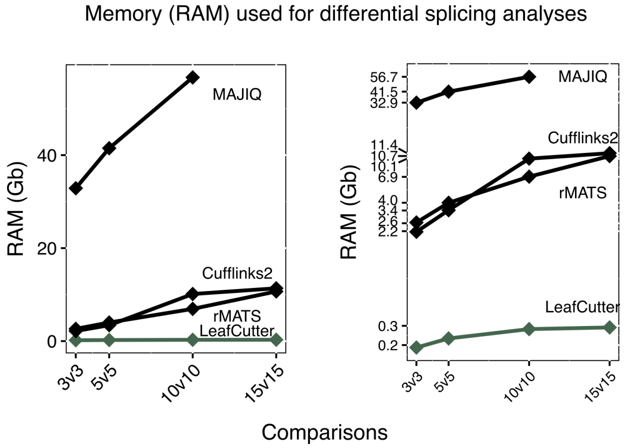
Supplementary Figure 3: **(a)** Distribution of the number of different GTEx tissues in which junctions predicted to be absent, or present in three commonly-used annotation databases, could be detected. **(b)** Relative junction usage in multiple GTEx organs of annotated and unannotated junctions identified in four GTEx organs. **(c)** Distribution of LeafCutter clusters from GTEx samples in terms of their splicing types. Clusters with only annotated junctions and clusters with unannotated junctions were further separated.



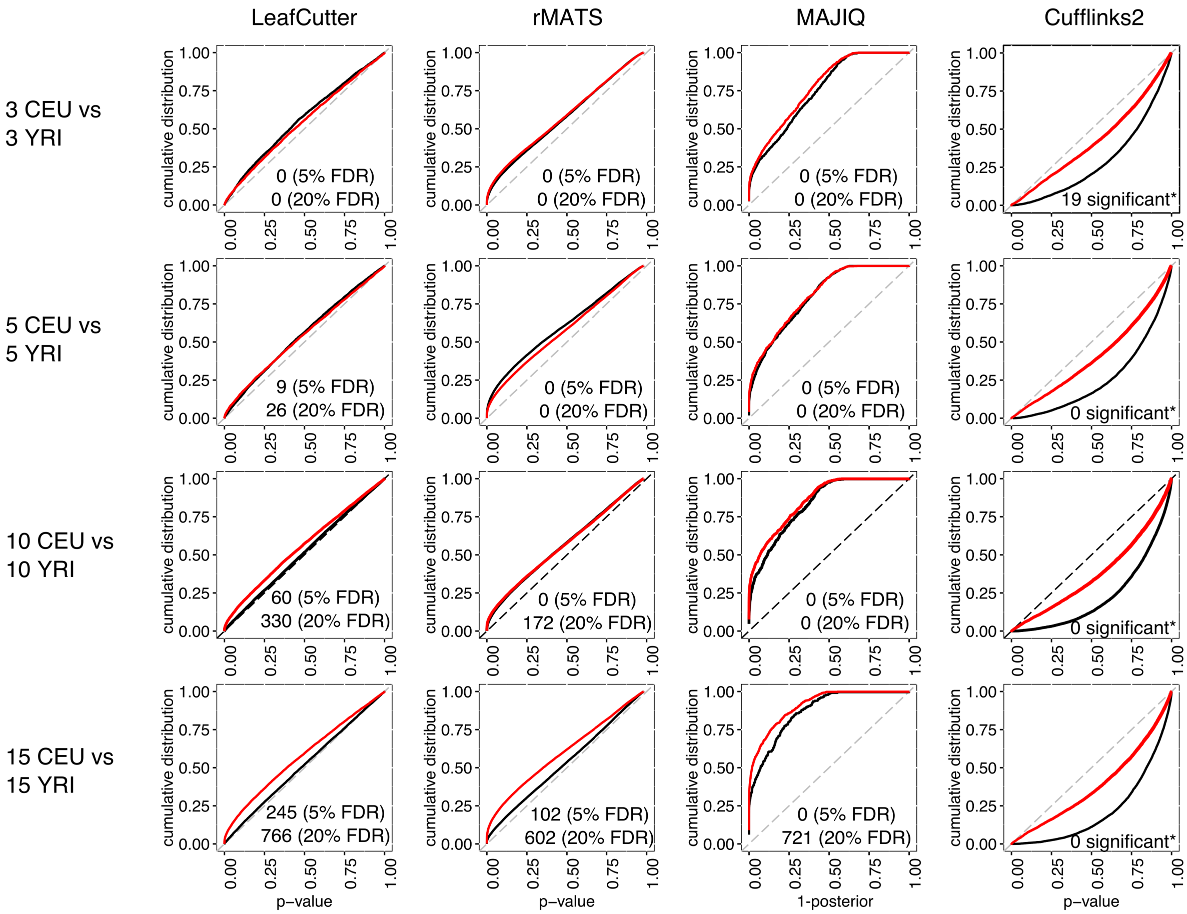
Supplementary Figure 4: PhastCons score distribution of splice site of novel introns. While ∼60% of annotated splice sites have local phastCons score >0.6, only 15-25% of unannotated splice sites do. Thus ∼80% of novel splice sites may represent noisy intron excision events.



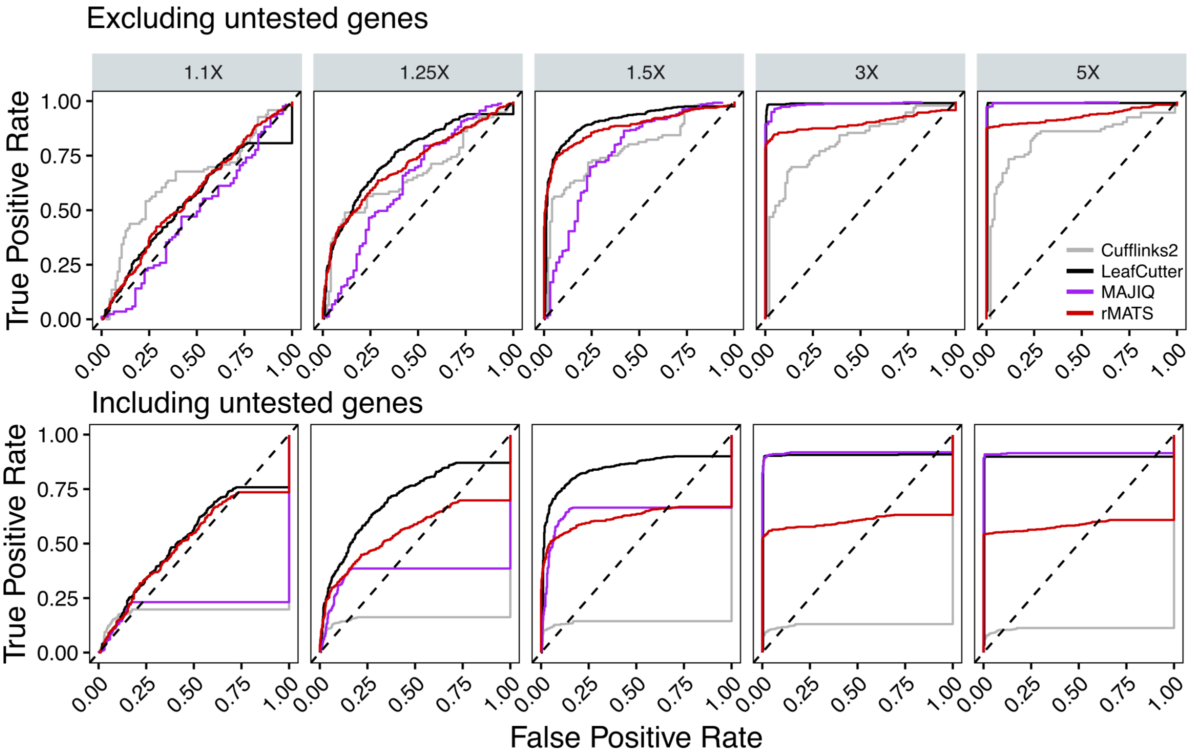
Supplementary Figure 5: Comparison between beta-binomial and Dirichlet-multinomial models for differential splicing analyses, performed on 10 male brain vs. heart samples from GTEx. Two approaches for combining per-intron p-values into cluster level introns are compared: Bonferroni correction and Fisher's combined test. Bonferroni is very conservative, as expected. Fisher's combined test has considerably lower power than the multinomial approaches. However, only v2 of the Dirichlet-multinomial (which uses a per intron concentration/overdispersion parameter) is well calibrated under permutations.



Supplementary Figure 6: Memory usage (RAM) of four differential splicing methods applied to comparisons between 3, 5, 10, and 15 YRI vs CEU LCLs RNA-seq samples. We omitted the 15v15 MAJIQ run due to its expensive resource usage (both in terms of time and RAM). Right panel shows usage in log scale.

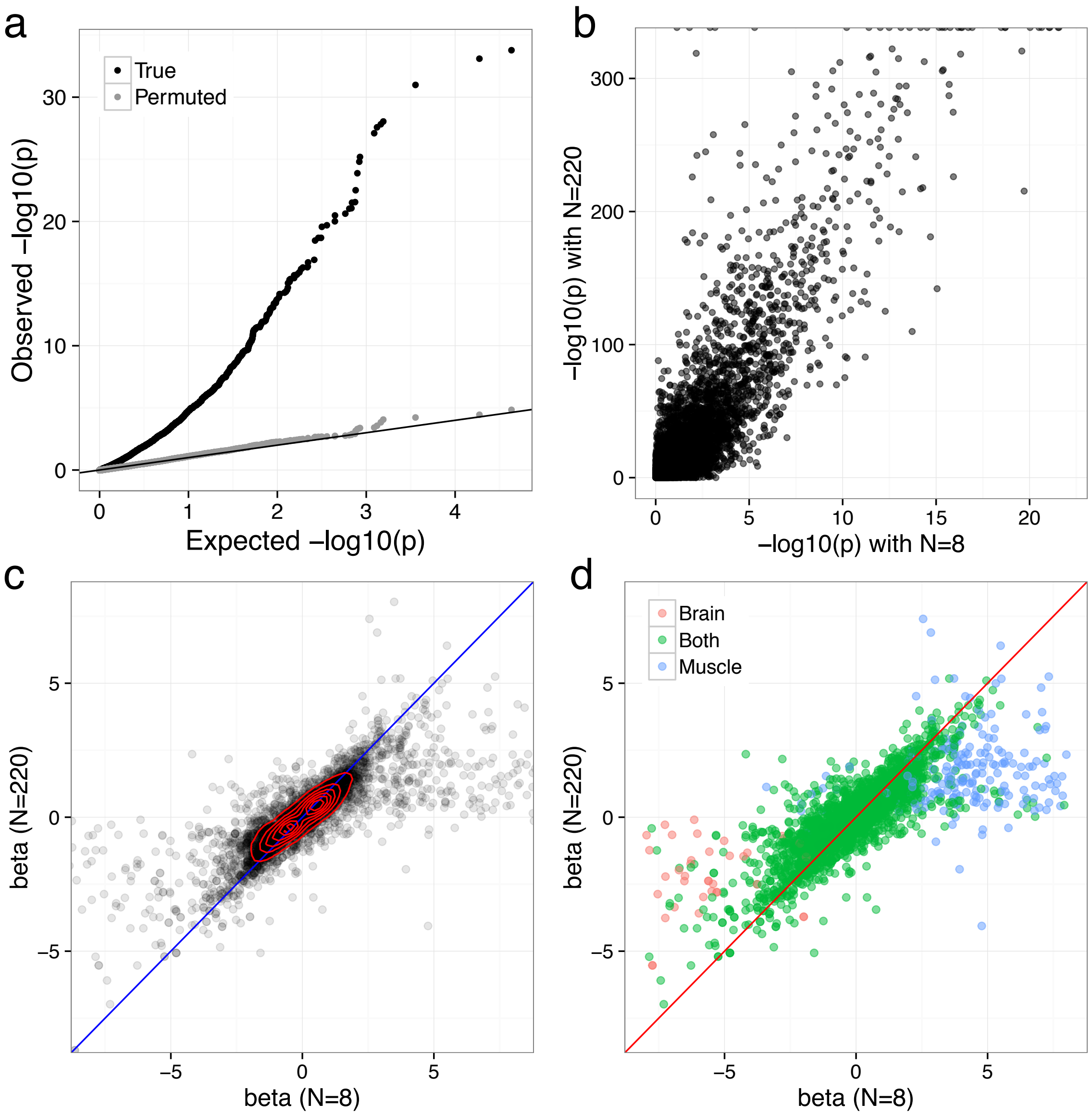


Supplementary Figure 7: Cumulative distributions of differential splicing test p-values (1-posterior for MAJIQ) for the all YRI versus CEU LCLs comparison (red). The distribution of test p-values for the permuted comparisons are also shown (black). * Cufflinks2 reports 19 significantly differentially spliced genes in the 3 vs 3 comparison, but none in the other comparisons.

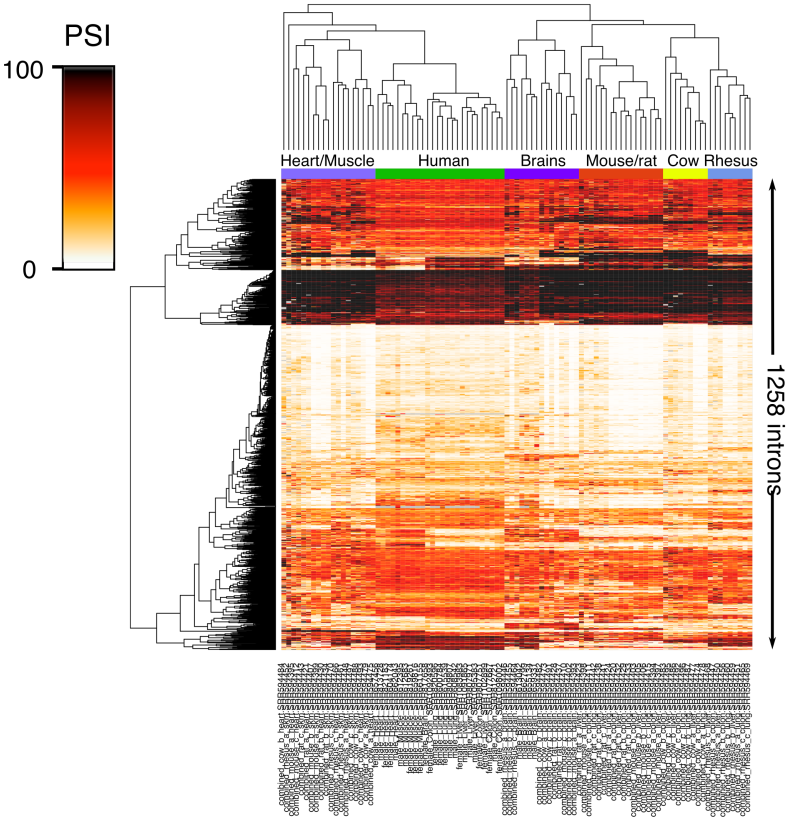


Supplementary Figure 8: Receiver operating characteristic (ROC) curve of LeafCutter, Cufflinks2, rMATS and MAJIQ when evaluating differential splicing of genes with transcripts simulated to have varying levels of differential expression. Top panel shows ROC curves when excluding genes that were not tested by each respective methods. While the bottom plot includes genes that were not tested in the calculation of true positive rate.

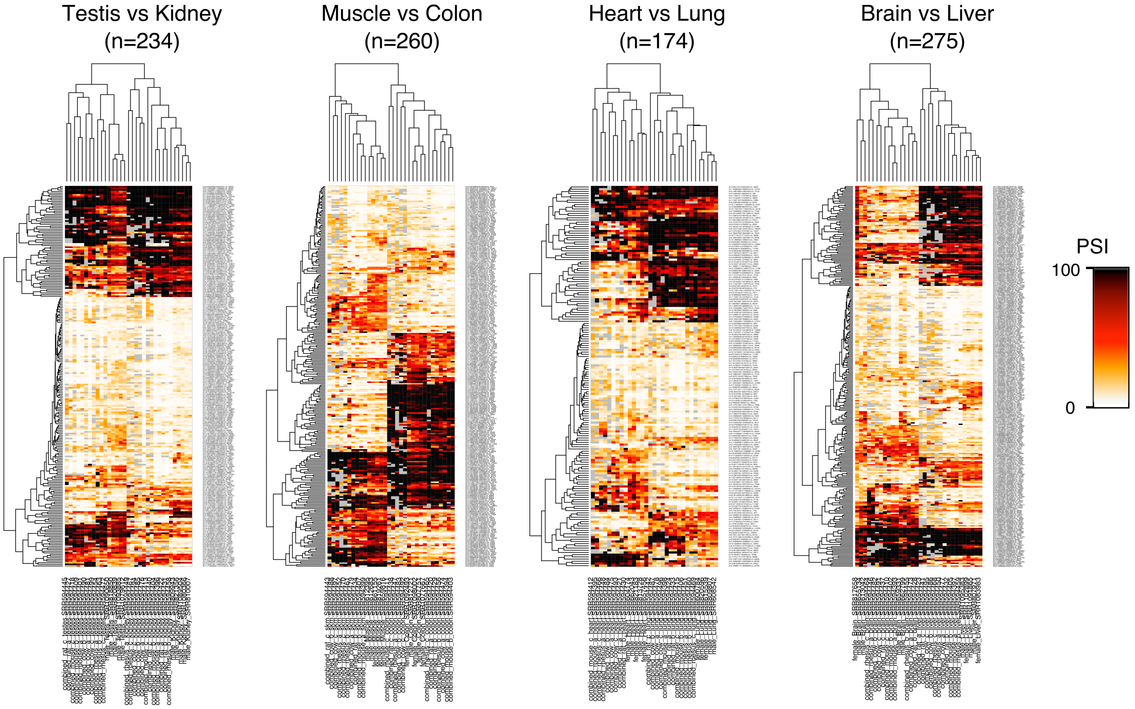
13 Supplementary Note Figures



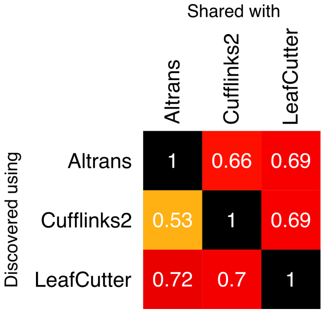
Supplementary Figure 9: LeafCutter is effective even with as few as 8 samples. Here we performed differential splicing analysis of 4 male brain vs 4 male muscle samples, and compared to results using 220 samples. **a)** p-values under permutations are well-calibrated. **b-c)** p-values and effect sizes are highly correlated between the two sample size datasets. **d)** Significant disparity in effect sizes between the two sample sizes is primarily driven by an intron being unique to a tissue when $N = 8$.



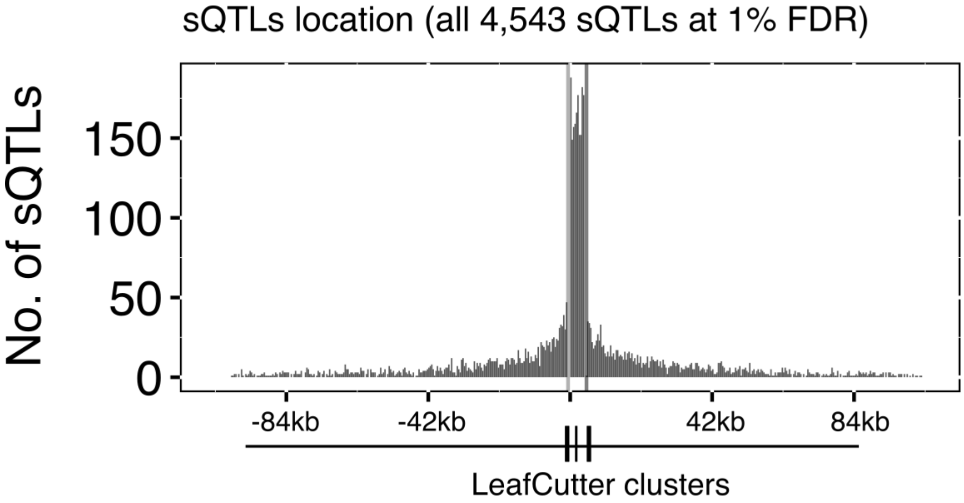
Supplementary Figure 10: Hierarchical clustering on all 1,258 introns that had no missing values in any of the samples.



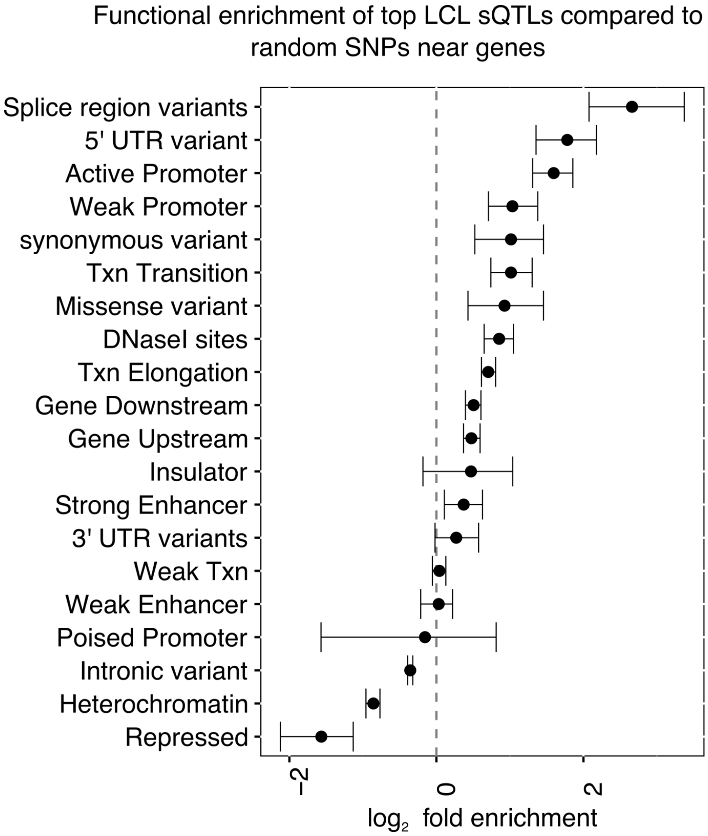
Supplementary Figure 11: We restricted to introns that were found to be differentially excised between human tissues (p-value $< 10^{-10}$ and effect size > 1.0)



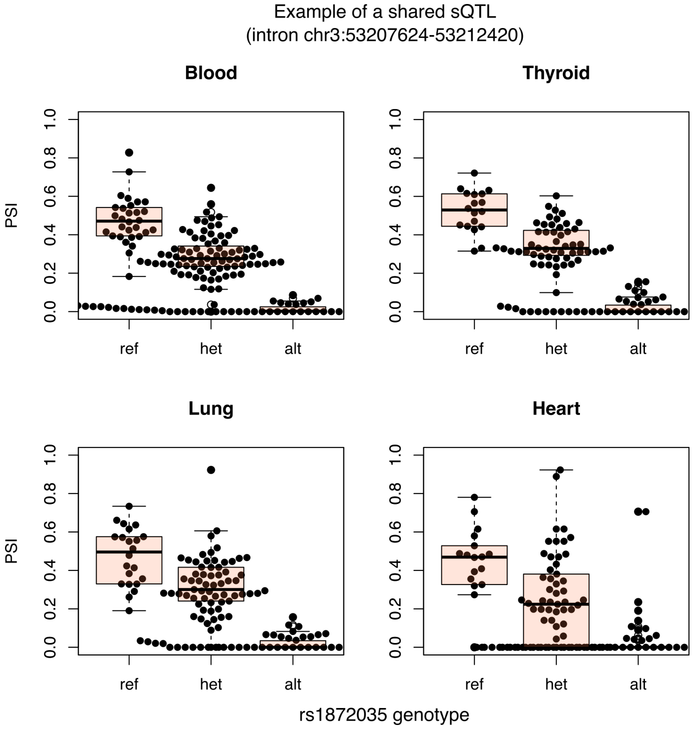
Supplementary Figure 12: Sharing of sQTL discoveries between Cufflinks2, Altrans, and LeafCutter estimated using Storey's π_0 method.



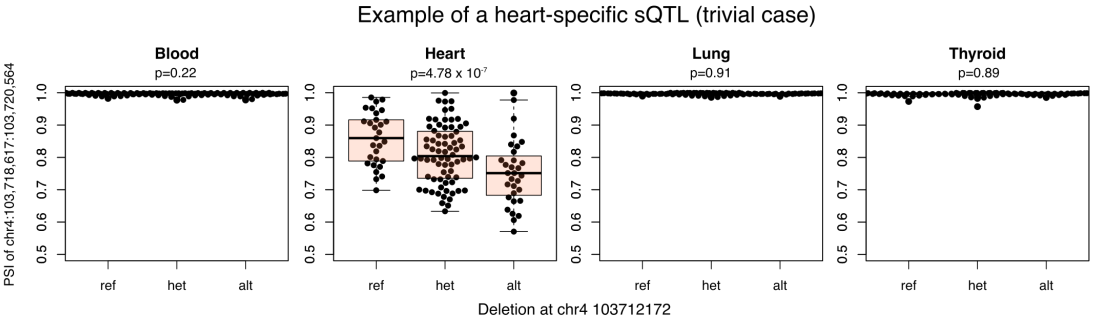
Supplementary Figure 13: Meta-cluster representation of position of all 4,543 sQTLs identified at 1%FDR.



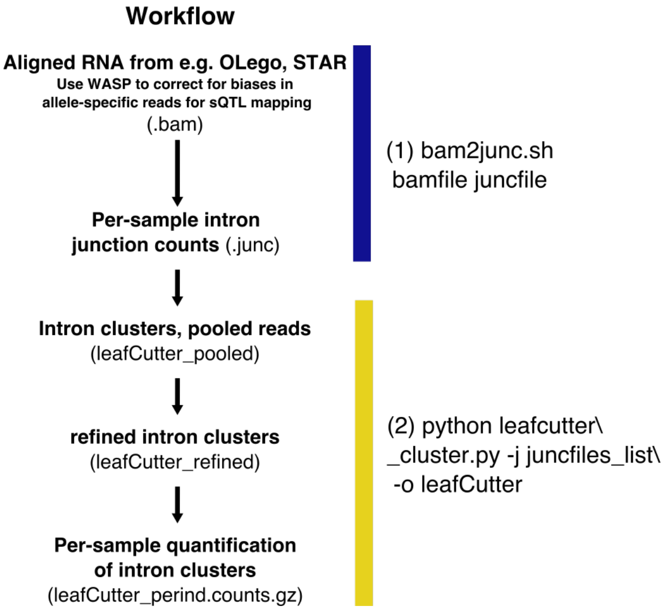
Supplementary Figure 14: Functional enrichment of 4,543 sQTLs identified at 1%FDR from CEU GEUVADIS data. Bar represent 95% confidence interval from 500 bootstraps.



Supplementary Figure 15: Example of a shared sQTL.



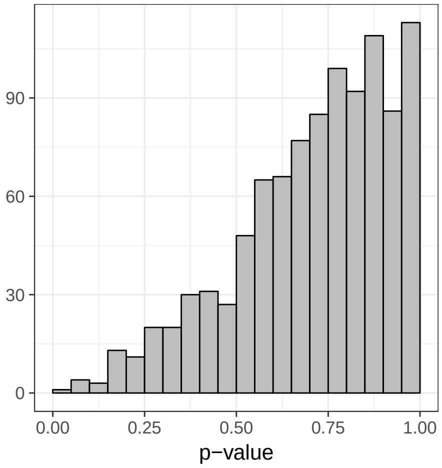
Supplementary Figure 16: Example of a tissue-specific sQTL.


Workflow
Aligned RNA from e.g. OLego, STAR
Use WASP to correct for biases in
allele-specific reads for sQTL mapping
(.bam)
Per-sample intron
junction counts (.junc)
Intron clusters, pooled reads
(leafCutter_pooled)
refined intron clusters
(leafCutter_refined)
Per-sample quantification
of intron clusters
(leafCutter_perind.counts.gz)

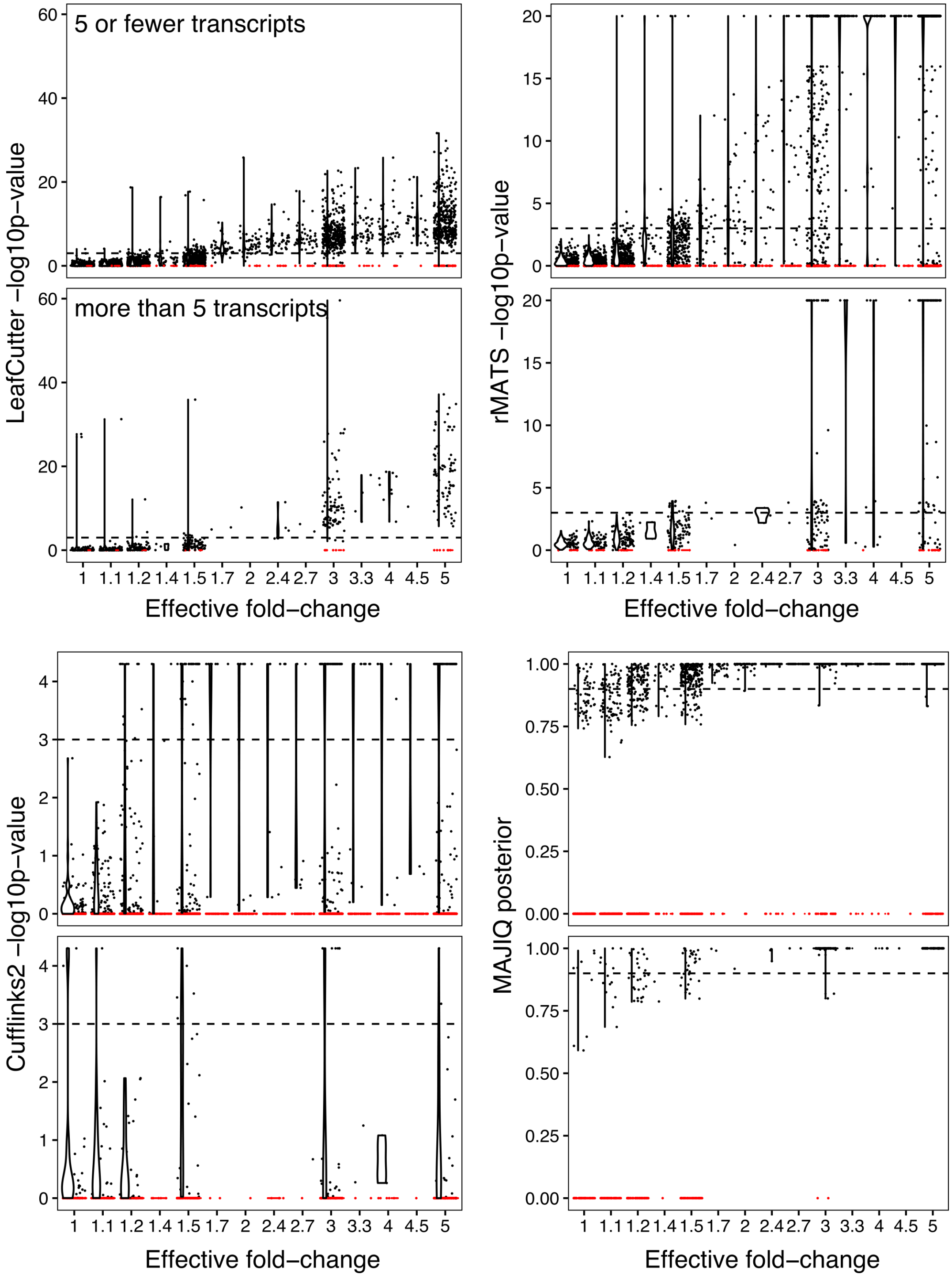
```
(1) bam2junc.sh
 bamfile juncfile
```

```
(2) python leafcutter\
_cluster.py -j juncfiles_list\
 -o leafCutter
```

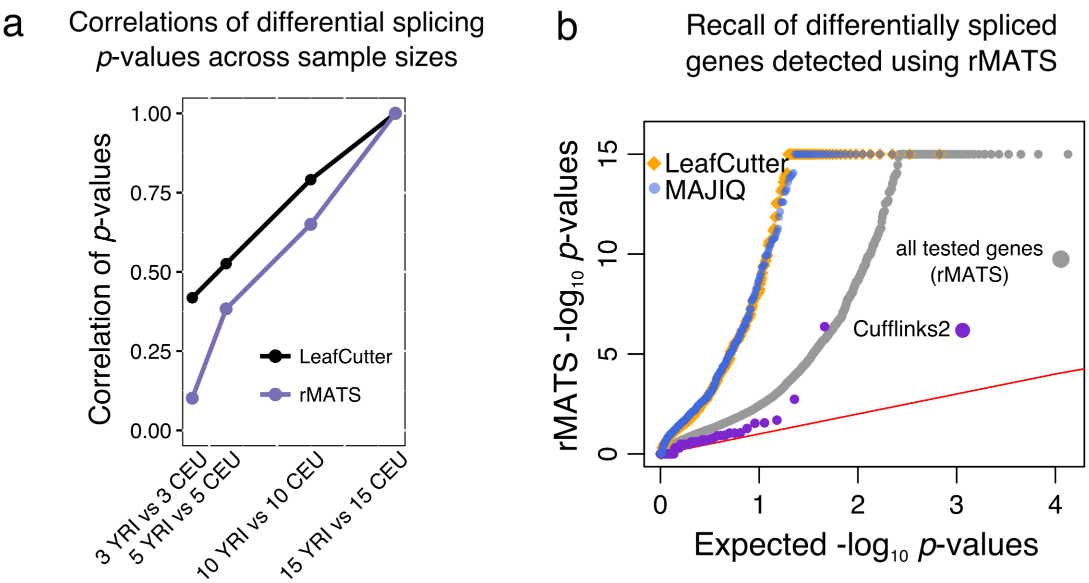
Supplementary Note Figure 1: Helper method and LeafCutter workflow for intron clustering.



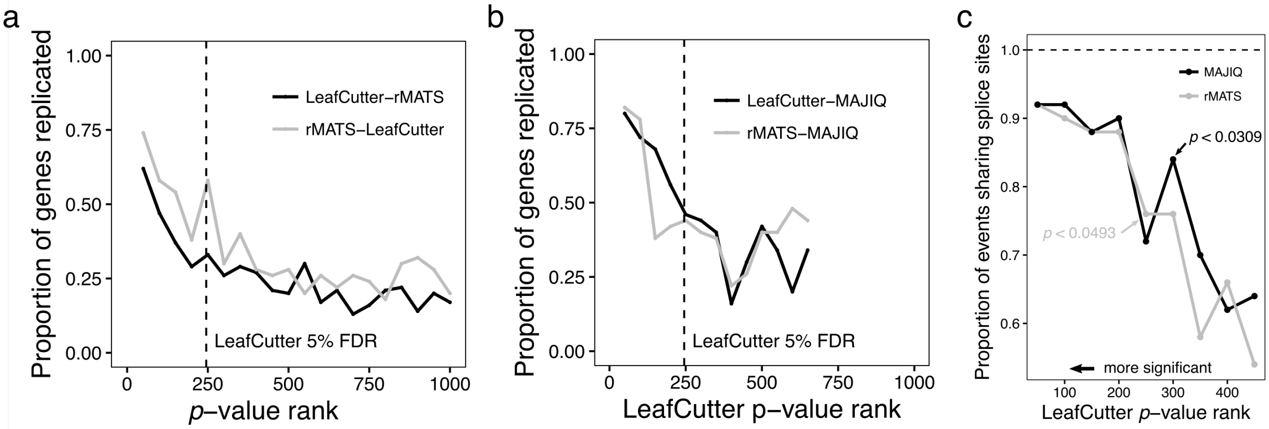
Supplementary Note Figure 2: Simulated isoform usage under the null of no differential splicing shows Cufflinks2 p-values are overly conservative.



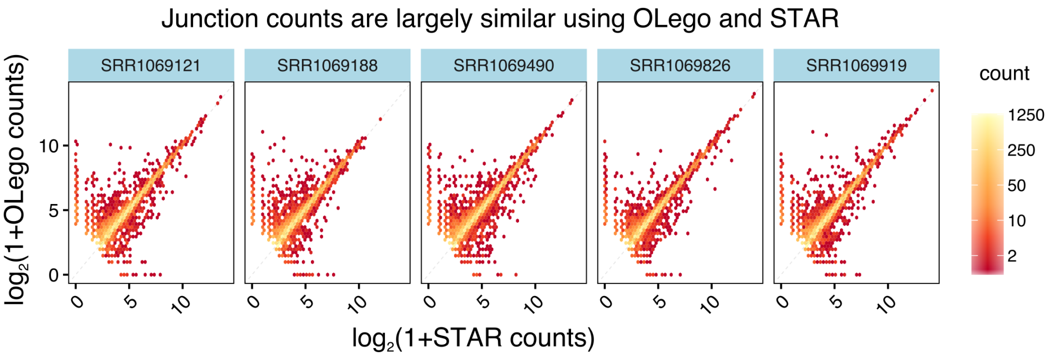
Supplementary Note Figure 3: Scatter and violin plots of the p-value and posterior distribution of differential test statistics binned by true, simulated, effective transcript fold-change. For each method, tests of genes with five or fewer transcripts and genes with more than five transcripts are plotted on the upper and bottom panels, respectively. We observed a decrease in power to detect differential splicing as transcript number increases using Cufflinks2, but not for the three other methods. Red dots represent genes with no tested splicing event.



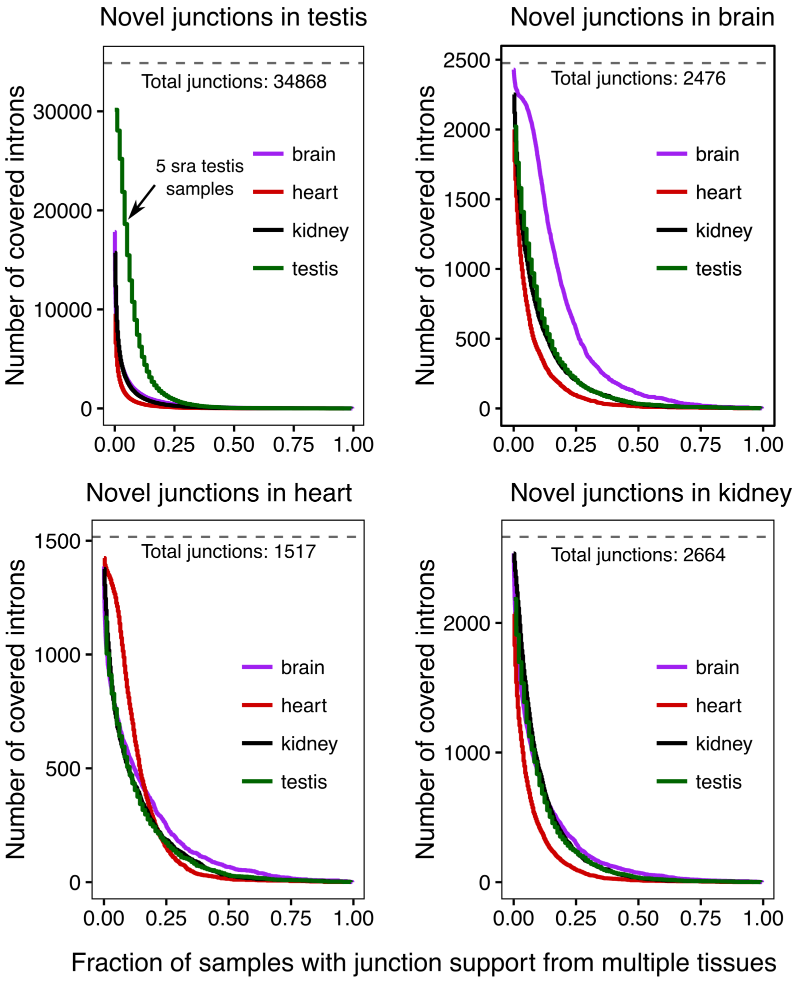
Supplementary Note Figure 4: **(a)** Correlation of computed differential splicing $-\log_{10}$ (p-values) of introns between a 15 YRI vs 15 CEU LCLs comparison and 3 vs 3, 5 vs 5, and 10 vs 10 comparisons. **(b)** QQ-plot of the differentially splicing signal found using rMATS in a comparison between 15 YRI and 15 CEU LCLs samples. Differentially spliced genes detected using LeafCutter and MAJIQ, but not Cufflinks2, are highly enriched in genes detected using rMATS.



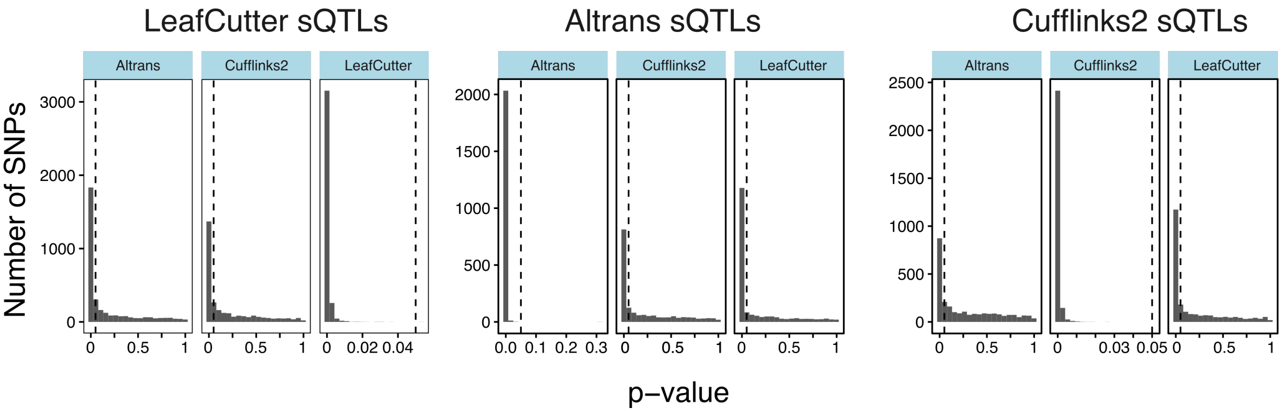
Supplementary Note Figure 5: Estimates of concordances between differentially spliced genes detected using LeafCutter and rMATS genes **(a)** and between LeafCutter or rMATS genes and MAJIQ genes **(b)**. Genes were ranked in terms of their significance levels (from LeafCutter and rMATS) and grouped into bins of size 50. Dashed lines mark 245, i.e. the number of differentially spliced genes detected using LeafCutter at 5% FDR. **(c)** Estimates of the proportion of shared splice sites between differentially spliced introns predicted using LeafCutter and introns predicted to be differentially spliced using rMATS and MAJIQ. Genes were ranked in terms of their significance levels (LeafCutter) and grouped into bins of size 50.



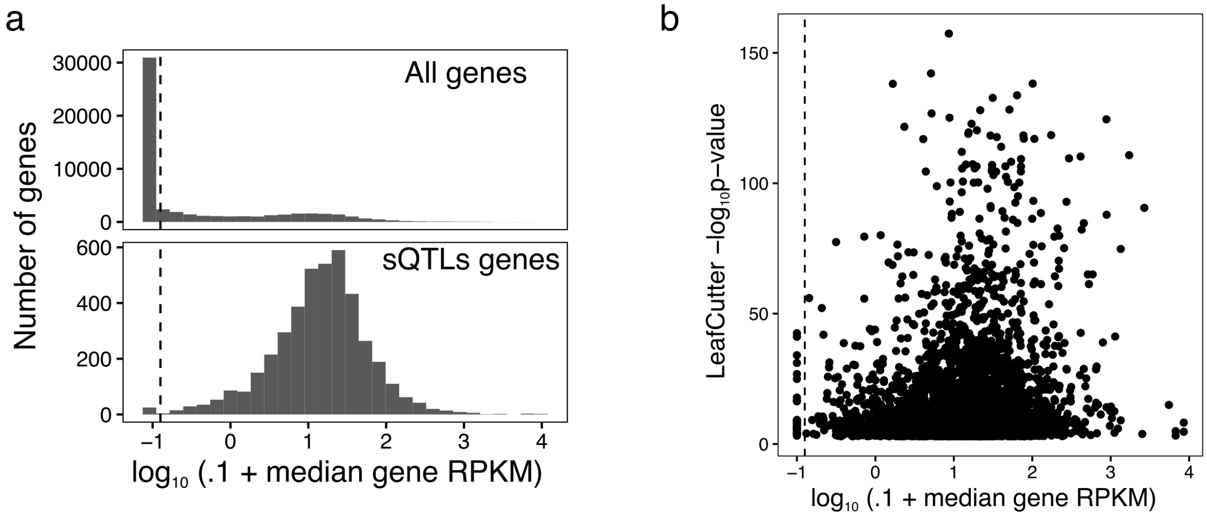
Supplementary Note Figure 6: Comparisons of junction read numbers between STAR and OLego across 5 random GTEx samples. Only junctions with total reads of more than 16, across both aligners, are shown. Note that only junctions which were found using OLego in a bigger panel of GTEx tissues (i.e. all GTEx samples in this study) were considered.



Supplementary Note Figure 7: Number of junctions that were found in at least X percent of all SRA samples, by tissue.



Supplementary Note Figure 8: Distribution of SNP-gene splicing association p-values. Three panels correspond to sQTLs identified at 10% FDR using LeafCutter, Altrans, and Cufflinks2, respectively.



Supplementary Note Figure 9: **(a)** Distribution of median LCLs gene expression levels for all genes (top) and genes with one or more LeafCutter sQTLs. **(b)** Scatter plot of LeafCutter *p*-value associations with respect to the expression levels of the corresponding genes. Dashed lines correspond to approximately 0.025 RPKM.

References

[1] Frazee, A. C., Jaffe, A. E., Langmead, B. & Leek, J. T. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* **31**, 2778–2784 (2015).

[2] Wu, J., Anczukow, O., Krainer, A. R., Zhang, M. Q. & Zhang, C. OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res.* **41**, 5149–5163 (2013).

[3] Li, Y. I., Sanchez-Pulido, L., Haerty, W. & Ponting, C. P. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Res.* **25**, 1–13 (2015).

[4] Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

[5] van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).

[6] Nellore, A. *et al.* Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biol.* **17**, 266 (2016).

[7] Nellore, A. *et al.* Rail-RNA: scalable analysis of RNA-seq splicing and coverage. *Bioinformatics* (2016).

[8] Ellis, S. E., Collado Torres, L. & Leek, J. Improving the value of public rna-seq expression data by phenotype prediction. *bioRxiv* (2017). arXiv:`http://www.biorxiv.org/content/early/2017/06/03/145656.full.pdf`.

[9] Carpenter, B. *et al.* Stan: a probabilistic programming language. *Journal of Statistical Software* **1** (2015).

[10] Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**, 1593–1599 (2012).

[11] Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).

[12] Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).

[13] Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9440–9445 (2003).

[14] Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).

[15] Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).

[16] Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

[17] Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* **47**, 1091–1098 (2015).

[18] Barbeira, A. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *bioRxiv* 045260 (2017).

[19] Wheeler, H. E. *et al.* Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. *PLoS Genet.* **12**, e1006423 (2016).